Memory, 2013
Vol. 21, No. 8, 945–968, http://dx.doi.org/10.1080/09658211.2013.769607

Routledge
Taylor & Francis Group

Retention-error patterns in complex alphanumeric serial-recall tasks

Fabien Mathy[1] **and Jean-Stéphane Varré**[2]

[1]Université de Franche-Comté, France
[2]Université de Lille, France

We propose a new method based on an algorithm usually dedicated to DNA sequence alignment in order to both reliably score short-term memory performance on immediate serial-recall tasks and analyse retention-error patterns. There can be considerable confusion on how performance on immediate serial list recall tasks is scored, especially when the to-be-remembered items are sampled with replacement. We discuss the utility of sequence-alignment algorithms to compare the stimuli to the participants' responses. The idea is that deletion, substitution, translocation, and insertion errors, which are typical in DNA, are also typical putative errors in short-term memory (respectively omission, confusion, permutation, and intrusion errors). We analyse four data sets in which alphanumeric lists included a few (or many) repetitions. After examining the method on two simple data sets, we show that sequence alignment offers 1) a compelling method for measuring capacity in terms of chunks when many regularities are introduced in the material (third data set) and 2) a reliable estimator of individual differences in short-term memory capacity. This study illustrates the difficulty of arriving at a good measure of short-term memory performance, and also attempts to characterise the primary factors underpinning remembering and forgetting.

Keywords: Memory; Span; Scoring; Errors; Short-term memory; Working memory; Chunks; Chunking.

The ability to recall a sequence of items in order is a fundamental psychological process that depends on many determinants, among them are: phonological features (e.g., articulation duration, phonological similarity, irrelevant sound; see Baddeley, 1986), temporal distinctiveness (Brown, Neath, & Chater, 2007), interference (Oberauer, Lewandowsky, Farrell, Jarrold, & Greaves, 2012), and decay (Barrouillet & Camos, 2012). Although some arguments can be advanced to support the idea that there exists no short-term or working memory functionally distinct from long-term memory (Surprenant & Neath, 2009), short-term memory and working memory usually refer to temporary storage and to temporary storage plus the processing of representations respectively (Aben, Stapert, & Blokland, 2012). Much of the research on immediate serial recall is most often associated with the concept of short-term memory, but because working memory tasks also require participants to immediately recall a list of items (albeit greater emphasis is

Address correspondence to: Fabien Mathy, Université de Franche-Comté - UFR SLHS, 30-32 rue Mégevand, 25030 Besançon Cedex, France. E-mail: fabien.mathy@univ-fcomte.fr

This research was supported in part by the Agence Nationale de la Recherche Grant # ANR-09-JCJC-0131-01 and in part by the Centre National de la Recherche Scientifique (CNRS) and the Institut National des Sciences Humaines et Sociales (INSHS) research grant (PEPS) to Fabien Mathy in 2009. The authors wish to express thanks to Nicolas Heller for providing valuable assistance with analyzing data.

This article was originally published with errors. This version has been corrected. Please see erratum (http://dx.doi.org/10.1080/09658211.2013.787811).

© 2013 Taylor & Francis

placed on processing demands), we refer alternatively to short-term memory and working memory tasks and processes throughout the present study. We aim at discussing retention-error patterns and capacity estimates in both domains.

Span tasks

The concepts of short-term memory and working memory are often used interchangeably (Aben et al., 2012), and the problem is further complicated by a large variety of tasks that do not necessarily follow a clear maintenance/(maintenance + manipulation) distinction. For instance, simple span tasks (e.g., digit span, letter span, word span) and complex span tasks (e.g., reading span, operation span) are assumed by many to measure short-term memory capacity and working memory capacity respectively (Conway et al., 2005; Conway, Kane, & Engle, 2003). However, we believe that it is not certain that these particular tasks strictly represent the *storage* versus *(storage + processing)* components. On the one hand, because simple span tasks allow the to-be-remembered material to be processed freely (e.g., some digits can be associated with one another to facilitate storage), they do not exclusively concern the storage of information. In spite of this lack of specification, short-term memory tasks (sometimes called in a more neutral fashion immediate-serial-recall tasks) appear rather seductive because of their simplicity and are hence considered as touchstone tasks for testing, for instance, whether memorisation is based on associations (Lewandowsky & Murdock, 1989; Murdock, 1995), positions (Burgess & Hitch, 1999; Henson, 1998b) or order (Page & Norris, 1998) —see Henson (1998b) for a review of these three theories of how a novel sequence of items is stored and retrieved in the correct order.

On the other hand, many complex span tasks involve a concurrent processing task (Conway et al., 2003; Lewandowsky, Oberauer, Yang, & Ecker, 2010), which is hypothesised to permit a separation of the storage and the processing components. In such tasks, the storage of the to-be-remembered material alternates with the processing of the not-to-be-remembered material (since by design, the concurrent task directs the processing component away from the maintained items). Hence, the processing component is dismissed rather than investigated. We believe that the measures within simple and complex span tasks evolved toward the following distinction: *storage + limited processing* for short-term memory span tasks and *storage* for many complex span tasks (although other complex span tasks involve the simultaneous processing and storage of the maintained items, e.g., the backwards digit span, which requires one to immediately recall a sequence of digits in the reverse order, see de Abreu, Conway, & Gathercole, 2010; and, possibly, the running memory span, see Bunting, Cowan, & Saults, 2006; Morris & Jones, 1990). As a result, the frequent entanglement of the different tasks (and the components they are supposed to reflect) might explain why sometimes there is a similar contribution of simple and complex span tasks to the prediction of intelligence (e.g., Colom, Rebollo, Abad, & Shih, 2006; Engle, 2002; Unsworth & Engle, 2006, 2007; Unsworth, Redick, Heitz, Broadway, & Engle, 2009).

Span tasks including repetitions

A to-be-tested idea is that the processing component can play a more major role when dedicated to maintaining the to-be-remembered items. The storage-processing combination might be the most correct composite variable for predicting cognitive abilities, but mostly when the stored items are also processed. This is an issue that we recently started to tackle (Chekaf & Mathy, 2012, in revision; Mathy & Feldman, 2009), by developing the idea that the role of the processing component can be enhanced when inviting participants to chunk the items to be stored. In this framework, the processing of the maintained items is induced by the use of repetition schemas in the material, which favour a recoding process (see also Fischer-Baum, 2012, where repetitions were also used to reveal the rich ways in which participants process sequential information). However, as demonstrated in the present study, the use of repetitions within lists of items requires a specific method to score memory performance reliably. This paper provides an introduction to this methodology.

Since the study by Miller (1956), most efforts in studying short-term memory capacity have focused on obtaining a pure measure of capacity limits by blocking the recoding of the to-be-remembered material. For instance, because the set of letters “b-h-n” is less meaningful than “u-s-a”, it is believed that people need to form three independent representations to retain “b-h-n”, whereas only one is required for ‘‘u-s-a’’. The recoding of ‘‘u-s-a’’ in ‘‘usa’’ is an example of an unwanted artefact in short-term memory experiments. As a result, a series of memory capacity tasks (usually referred to as working memory tasks) have been developed using more controlled material to prohibit recoding. The current project takes an opposite direction, by analysing individuals’ errors in non conventional serial-recall tasks in which the alphanumeric stimuli are sampled with replacement. Such tasks can generate more interference and/or facilitate the grouping of items, and as a result are more complicated to score. Immediate serial-recall tasks and simple span tasks can be synonymous, but since our serial-recall tasks cannot be considered simple, the second terminology (i.e., simple span tasks) is avoided in this paper.

Scoring methods

Scoring is fundamental, but the choice of scoring procedures can considerably change the estimates for a given span task (see Conway et al., 2005, pp. 774–775, who compare four basing scoring procedures; some examples are given by Cowan, 2001, p. 100; see also Martin, 1978, in the context of immediate free recall; St Clair-Thompson & Sykes, 2010; Unsworth & Engle, 2007) and can change reliability (Friedman & Miyake, 2005). For instance, some researchers advocate that there might not be such a fundamental difference between the magical number 4 (maybe reflecting a 100% correct performance criterion) and the magical number 7 (using a 50% criterion; see commentary by B. L. Bachelder in Cowan, 2001, p.116; Broadbent, 1975, also advised taking performance discontinuities into account). In most psychometric tasks, participants are presented with increasingly long series of elements. For instance, participants are presented with three series of each length, until they fail to recall the elements of all three series at a particular level. In that case, the correct recall of all the series of one, two, and three elements, of two series of four elements and one series of five elements can result in a span of $(3+3+3+2+1)\times 1/3=4$ (Barrouillet, Bernardin, & Camos, 2004, p. 88; Conway et al., 2005’s all-or-nothing scoring) or can result in a span of $(3+3+3+(2+.75)+(1+.6+.6)\times 1/3=4.65$ if one takes into account the number of elements correctly recalled in each series (the participant recalled in that case once 3 elements out of 4, and twice 3 elements out of 5; see Conway et al., 2005’s partial-credit scoring). Our objective is to demonstrate that scoring supraspan lists (i.e., the lists that exceed capacity) can be facilitated by sequence alignment methods. Such methods present the additional benefit of determining the errors individuals make when they misrecall a list.

Examples of studies interested in supraspan conditions include those which evaluate the retention span by integrating the area under the serial-position curves (Brown et al., 2007; Murdock, 1962). However, again, different scoring criteria can lead to substantial differences in the serial-position curves obtained: for instance, Farrell and Lewandowsky (2002) showed how two serial-position curves differed when either a strict criterion or a less stringent criterion was used (p. 10). To take another example of how scoring can prove tricky, Henson (1998b) recommended scoring by input position or by output position (p. 124), but acknowledged that the distribution of omissions plotted against input position can differ from that plotted against output position. As a last example, McCormack, Brown, Vousden, and Henson (2000) pointed out that because there were inherent ambiguities in the scoring method they employed in their task, they had to use lists of the same length to facilitate comparisons and a written serial-recall procedure that enabled them to classify most errors unambiguously. None of these procedures are considered standard, straightforward, and generalisable. To make matters worse, these procedures are simply inadequate to score the lists that include repetition. Such material unconditionally requires other methods in order to compute partial-credit scoring.

The method that we develop in this study is unfortunately —in its current development— no more standard, straightforward, and generalisable than those cited above. However, we believe that the idea of developing a method for lists that include repetition is not a straw man argument because the literature shows much promise in studying learning material that present repeatable patterns for which similar methodological difficulties might arise (e.g., grammar learning and chunking: French, Addyman, & Mareschal, 2011; Servan-Schreiber & Anderson, 1990; short-term memory: Della Sala, Gray, Baddeley, Allamano, & Wilson, 1999; Henson, 1998a; Mathy & Feldman, 2012). There is effectively an obvious need to study lists with repetition in order to study interference processes and chunking abilities. For instance, by mixing items of dissimilar kinds (e.g., 1564782 instead of repeatable items 11001011010101), the possibilities for studying interference and chunking are rather limited. In our case, the method allows us to make an estimate of the amount of information individuals can hold in short-term memory in various tasks. Specifically, we target an estimation of capacity irrespective of response accuracy, that is, the actual number of items or chunks that can be correctly recalled by participants immediately after presentation. Again, it must be pointed out that we do not intend to introduce a new golden standard method for computing memory capacity since the method presents several options that can prevent its standardisation. However, there is a practical difficulty of scoring performance when repetition or regularity are introduced in the material, and we believe that the use of sequence alignment algorithms is the best option, if not the only one, to obtain a reliable estimation of the material recalled in the four experiments presented in this paper.

We now develop a short argument against studying oversimplified lists. Although it simplifies the scoring process, experimenting with lists of items sampled without replacement has disadvantages as compared to confusable tasks (e.g., those in which items are drawn with replacement). When digits are used, for instance, conditional probabilities are not the same across the sequence. Suppose the stimulus is 124896537, and the participant can only recall the first eight digits 12489653, there is still a $p = 1$ chance that 7 is the last digit to be remembered, so it will be recalled more easily by a participant for whom the span is 8 items. In other words, p simply increases with list length (see Martin, 1978, Exp. 1; likewise for an experiment by Page & Norris, 1998, in which only a set of eight nonconfusable letters was used, with no repeats).

However, performance scoring can prove unreliable when the items to be recalled are drawn with replacement. For instance, given a *1100110* response for a 110010110 stimulus, we need to determine which portion was forgotten by the participant (e.g., 11001*0*110, or 110010*1*10, both of which lead to a *1100110* response). A simpler example would be a *125417* response to a 1251417 stimulus, for which a correct alignment of the items would clearly indicate that the item in fourth position was forgotten. A correct scoring of such stimuli might help estimate whether repetition allows the items to be less recalled when far apart, a phenomenon known as the Ranschburg effect (scoring difficulty in this context was identified by Henson, 1998a). The solution that we propose, namely aligning the stimulus and the response, will typically focus on both memory for the temporal occurrence of events and item memory, two highly debated processes (Anderson & Matessa, 1997;Botvinick & Plaut, 2006; Brown, Preece, & Hulme, 2000; Burgess & Hitch, 1999; Estes, 1997; Gallistel, 1990; Henson, 1998b; Henson, Norris, Page, & Baddeley, 1996; Lewandowsky & Murdock, 1989; Page & Norris, 1998).

To more reliably score short-term memory performance, we propose a new method based on an algorithm that is usually devoted to DNA sequence alignment. We discuss the utility of aligning the stimulus and the participant's response to get the most reliable pattern of errors. Determining the similarity between two sequences is a common task in computational biology. For instance, alignments are usually used to analyse phylogenetic relatedness between two DNA sequences. Our method does not depend on which computational model of short-term memory is considered for further analysis. Our conception is based on the idea that typical DNA errors are also typical putative errors in short-term memory. Although not directly referring to DNA sequencing—an approach that has never been considered before to the best of our knowledge—, the classification of errors made earlier by Henson (1998b, p. 123) perfectly fits DNA errors, so there is great incentive to use bioinformatic tools. Errors can be categorised as follows: deletion/omission (forgetting), substitution/confusion (an item is replaced by another), translocation/transposition/movement errors (an item is recalled in an incorrect position), and insertion/intrusion errors (an item is inserted during recall, or an item is erroneously repeated). The last type of error occurs very rarely (Henson et al., 1996), which provides support for response suppression during recall (Lewandowsky, 1999). In certain commercial software packages, such algorithms are available in the form of simple functions that are almost as easy to run as a function for computing a mean (e.g., nwalign MATLAB® function, MATLAB®, Bioinformatics Toolbox™, The MathWorks Inc., Massachusetts). We used such algorithms to analyse a few new empirical datasets, for digits and letters drawn with replacement. For example, a benefit of the new scoring method is to go beyond the usual partial-credit scoring methods by better rectifying the span estimates when errors are committed by participants. Usual partial-credit scoring methods are vulnerable to differences in input position and output position of the items. For instance, when a participant makes an omission error on the first item, then all of the later items in the sequence would not be scored as correct. Our method makes use of an algorithm that aligns the recalled items to their original position to better credit the participant's memorisation.

This study illustrates the difficulty of arriving at a good measure of short-term memory performance, but also attempts to determine the primary factors underpinning remembering and forgetting. By applying a new scoring method, our principal goals are to estimate the span, to decompose memory errors by category, to calculate the number of errors by category, to estimate the number of chunks that can be encoded in short-term time, and to compare the spans estimated by different scoring methods. The first two experiments were carried out to evaluate the short-term memory span based on sequence alignment and to analyse retention-error patterns in simple-span tasks that required immediate serial recall of lists with possible repeated items. These two experiments were also conducted to investigate the proportion of deletion, substitution, and insertion errors by item position, the increase in substitution, insertion, and deletion errors with list length, and the number of items recalled as a function of list length. As demonstrated below, the advantage of the new method is that the span and retention-error patterns are estimated simultaneously in a single pass.

The aims of the last two experiments were twofold: (a) to evaluate the memory span for chunks based on sequence alignment (i.e., the number of chunks encoded given the number of chunks in a list) in immediate serial recall tasks; chunking processes were induced by introducing regularity in the lists; (b) to evaluate the relationship between the memory span for chunks and the span estimated by a common working memory battery (a question was whether chunking processes can adequately measure the processing component of working memory) and to show that low versus high working memory span groups differ in terms of retention-error patterns.

Experiment 1 and Experiment 2 respectively introduce a few repetitions in simple digit and letter span tasks in order to start applying our method to simple cases. We first run a basic analysis of performance for the first two data sets in the Results sections before computing the sequence alignments. We obtained reliable error frequencies by position, and a separate estimate of the increase in error rate with list length. We also show that the number of errors grows exponentially with list length (especially deletions and substitutions), that insertions and translocations are the rarest kind of errors, and that six items is the upper limit of the number of correctly recalled items in supraspan conditions (i.e., adding items beyond the person's memory span does not generate much more interference).

Experiment 3 and Experiment 4 focus on chunking memory span tasks. In Experiment 3, a data set from a previous publication (Mathy & Feldman, 2012) is reported. This data set was analysed here in greater detail in order to prove the benefit of having a correct stimulus-response alignment when computing the number of chunks that could be adopted by the participants (i.e., the actual number of correctly encoded and correctly recalled chunks, irrespective of response accuracy). This analysis also shows that when regularity is introduced in the lists, the actual number of chunks recalled by participants is asymptotic to four (independently of the number of unpacked items), which confirms previous observations (Cowan, 2001). Experiment 4 tested memory performance for lists that included several repetition ratios (from no repetition to random alternation of two items). Our motivation was that with more repetition, the information held in memory is further processed, in contrast to dual-tasks in which much effort is made by the experimenter to separate the maintenance and processing components (Baddeley & Hitch, 1974). In this last experiment, we confronted three scoring methods (all-or-nothing, partial-credit, and alignment) in the prediction of four working memory tasks developed by Lewandowsky et al. (2010). Our results show that performance for lists with repetitions computed with our alignment method offers the best estimator of individual differences in working memory capacity when the processing component is involved. All the results are discussed in the General Discussion section.

EXPERIMENT 1: RANDOM SEQUENCES OF DIGITS

Method

Participants. The participants were 37 students at the Université de Franche-Comté, France. They received course credit in exchange for their participation.

Procedure. In this computer-run experiment, participants were given an immediate serial-recall task. Each stimulus list of digits (chosen amongst the 0 to 9 range) was composed of at most 10 digits. On each trial, the entire list was presented sequentially to the participant at a pace of 1 digit per second (1-sec onset-to-onset time between digits). The participant was asked to immediately recall as many digits as possible in the order in which they were presented. The length of the list was random (from 3 to 10), rather than progressively increasing, to avoid confounding fatigue or learning effects with task difficulty effects (and to avoid other peculiar effects; see Conway et al., 2005, p. 773).

Each experimental session was limited to half an hour and included 100 separate stimulus lists. The 100 lists and their order were randomly drawn for each participant. The digits were randomly drawn with replacement (except that a digit could not occur twice in succession). In a given list of digits, each digit replaced the previous one in the same spatial location. After the presentation of the last digit of a given list, participants entered their response on a keyboard. The time for recall was unlimited and the participants were allowed to correct their response. They were asked to recall as many digits as possible in the correct order.

The digits entered by the participants were displayed on the computer screen (1 cm wide and 1.5 cm tall Arial letters) placed side by side to form a single row going from the participant's left to his/her right. The participants could read each response to make sure the sequence of digits they entered was what they had intended. Once their response was confirmed by pressing the space bar, the next list was shown.

Stimuli. The stimuli were displayed visually on the computer screen. Each stimulus was about 3 cm tall and was presented in the middle of the screen in white Arial font against a black background.

Results

The data from all participants were included in the analysis. The total number of digit sequences recalled by the participants was 3699.[1] The proportion of correct responses as a function of the number of digits is shown in Table 1 below.

Table 1 also shows that the participants' recollection of the stimuli was below 50% at around 7 digits ($p = .42$ at 7 digits). The odds ratio, which quantifies the discontinuity observed between 6 and 7 digits, was (326/458)/(189/451) = 1.7, meaning that the proportion correct was 1.7 times better for 6-digit stimuli than for 7-digit stimuli.

Figure 1 below shows the mean proportion of correct responses averaged across participants. Like Crannell and Parrish (1957) and many subsequent studies, we found an S-shaped function. The mean proportions are given in the last row of Table 1 below. A simple estimation of memory span from the sum of the proportion of correct responses across conditions (i.e., integrating under the performance curve) gave 6.2 (assuming that the proportion correct was 1 for both 1- and 2-digit sequences, since there were no such sequences in the data).

In order to detect the presence of discontinuities in the results, we ran several *t*-tests with repeated measures between adjacent conditions (i.e., 1 digit vs. 2 digits, 2 vs. 3, 3 vs. 4, etc.), adjusting the alpha level with Bonferroni's correction. All of the comparisons were significant except the one between 9 and 10 digits, but the greatest difference occurred between 6 and 7 digits ($M_d = .35$, $SD_d = .21$, $t(36) = 9.9$, $p < .001$), recalling the magical number seven capacity limit. A similar result was obtained by running a piecewise linear regression analysis to test the presence of a dramatic change occurring between 6 and 7 digits (6.5 was subtracted from the data in order to test the difference between the two intercepts at point 6.5 and separate slopes before and after 6.5). The multiple regression analysis could show a significant jump of .25 at 6.5 digits (although this was not the only significant breakpoint).

[1] One of our participants did not have sufficient time to finish the experiment. Because there was no indication of the number of lists completed on the screen, the experimenter could not know that the participant was one trial short of finishing the experiment.

TABLE 1

Correct-response data, as a function of the number of digits or the number of letters in the stimulus in Exp. 1 and Exp. 2

	Experiment 1 *Number of digits in the stimulus*							
	3	*4*	*5*	*6*	*7*	*8*	*9*	*10*
No. of correct responses	343	425	381	326	189	101	42	21
Total no. of responses	353	458	455	458	451	498	573	450
Proportion of correct responses	.97	.93	.84	.72	.42	.20	.07	.05
Mean of per-subject averages	.97	.92	.83	.73	.39	.26	.07	.05
	Experiment 2 *Number of letters in the stimulus*							
	3	*4*	*5*	*6*	*7*	*8*	*9*	*10*
No. of correct responses	337	279	227	235	83	49	49	15
Total no. of responses	347	318	294	397	280	321	439	344
Proportion of correct responses	.97	.88	.77	.59	.30	.15	.11	.04
Mean of per-subject averages	.96	.87	.78	.56	.29	.16	.09	.04

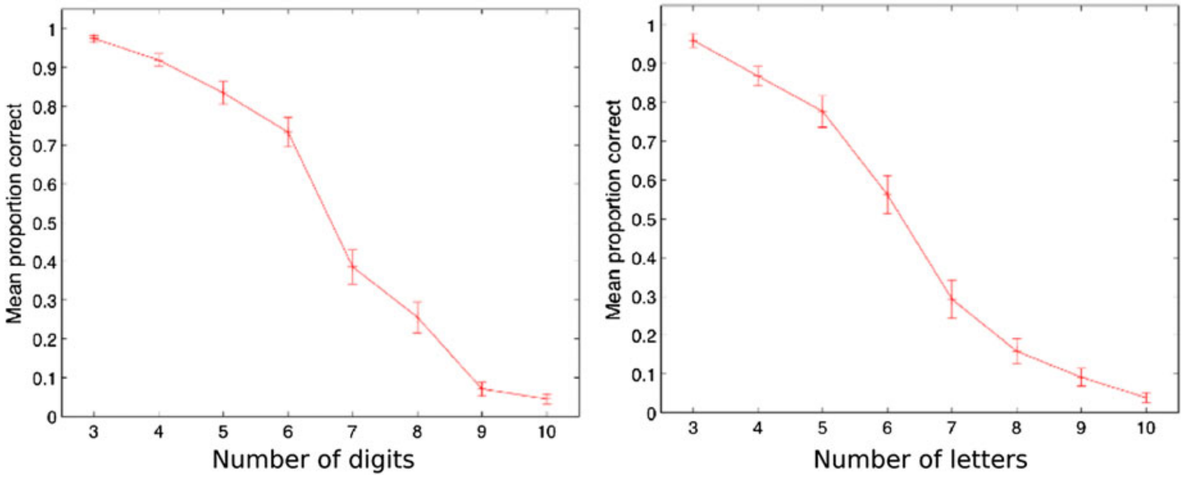

Figure 1. Mean proportion of correct responses as a function of the number of digits or the number of letters present in the stimulus list in Exp. 1 and Exp. 2. The mean of per-participant averages was calculated by averaging the proportions obtained for each participant. Error bars indicate ±one standard error.

EXPERIMENT 2: RANDOM SEQUENCES OF LETTERS

Method

Participants. The participants were 31 Franche-Comté University students who received course credit in exchange for their participation. None of them had participated in Experiment 1.

Procedure. The procedure was the same as in Experiment 1, except that list of letters were randomly drawn (with replacement) instead of digits.

Results

The data of all participants were included in the analysis. The total number of letter sequences recalled by the participants was 2740 (instead of the 3100 total that would have been obtained had the 31 participants been given more than half an hour to complete the 100 trials). The proportion of correct responses as a function of the number of letters is shown in Table 1 above.

Table 1 also shows that the participants' recollection of the stimuli was below below 50% around 7 letters ($p = .30$ at 7 letters). The odds ratio, which quantifies the discontinuity observed between

6 and 7 letters, was (235/397)/(83/280) $=2$, meaning that the proportion correct was twice as good for 6-letter stimuli as it was for 7-letter stimuli.

Figure 1 above shows the mean proportion of correct responses averaged across participants. The mean proportions are given in the last row of Table 1. A simple estimation of memory span from the sum of the proportion of correct responses across conditions gave 5.8 (i.e., integrating under the performance curve, assuming that the proportion correct was 1 for both 1- and 2-letter sequences, since there were no such sequences in the data).

In order to analyse the discontinuities in the results, we ran several *t*-tests with repeated measures between adjacent conditions (i.e., 1 letter vs. 2 letters, 2 vs. 3, 3 vs. 4, etc.), adjusting the alpha level with Bonferroni's correction. All discontinuities were significant except the one between 4 and 5 letters ($p < .016$, which did not reach the alpha level after correction) and the one between 8 and 9 letters ($p < .011$), but the greatest difference, again, occurred between 6 and 7 letters ($M_d = .27$, $SD_d = .21$, $t(30) = 7.0$, $p < .001$). A similar result was obtained by running a piecewise linear regression analysis that revealed a significant jump of .20 at 6.5 letters (although, again, this was not the only significant breakpoint).

SCORING BASED ON SEQUENCE ALIGNMENT ALGORITHMS

Our aim here was to characterise the retention-error patterns using sequence alignment algorithms (SAAs). Recall accuracy can simply be determined by analysing the number and position of deletions (forgotten items), mutations (substituted items, e.g., caused by acoustic confusion, Conrad, 1964), and insertion errors (adding an item that was not present in the original stimulus, or repeating an item at a later position). The search for translocations (transfer of a sequence to another location) or inversions (an entire sequence is reversed) is generally beyond the scope of SAA because such movements usually only concern large portions of DNA. Given that the permutation rates seemed very low in the data sets when we tested other ad hoc methods, we focused first on a basic algorithm run with default parameters, in view of searching for deletions, mutations, and insertions only.

Since there is empirical evidence that permutations are likely to occur in short-term memory tasks (Brown et al., 2000; Healy, 1974; Henson, 1996), it is still possible to modify basic algorithms to make them include the search for permutations between items, as demonstrated later. Because items are more likely to be recalled in a position near their original position than in a more distant one (a tendency called "locality constraint", Henson et al., 1996; Nairne, 1992), we computed the mean number of permutation errors that occurred between adjacent items (i.e., movement errors with a move distance equal to 1, such as "*ab*" instead of "ba"). Please note that all the responses made by subjects are in italics. There are numerous other cases in the following pages. The probability that two adjacent items were permuted was very low for both datasets, that is, $p = .064$ for digits (i.e., 1 pair out of 15; see also Majerus, Poncelet, Elsen, & van der Linden, 2006, p. 859, in which a similar .06 proportion was found in their immediate serial recall tasks) and $p = .024$ for letters (i.e., 1 pair out of 42; these low probabilities match those obtained by Henson et al., 1996, p. 91). Permutation likelihood apparently corresponds to the number of items in each stimulus set, since .064/.024 $\approx$1/(10/26). However, the number of permutations roughly estimated by searching the pairs that were found to be permuted *without prealignment of the stimulus and the response* is a very unsure estimate, especially when items are drawn with replacement. For instance, for a stimulus "13145" and a response *1345*, too simple an algorithm would search for permutations of the following pairs: 13, 31, 14, and 45; the algorithm would erroneously find one permutation (31 changed into *13*). However, a basic alignment of the digits would indicate instead that the third digit was deleted while the first two digits and the last two digits were correctly recalled. Accordingly, the number of permutations found after aligning the stimulus-response pairs was only .023 for digits and .022 for letters.

In bioinformatics, sequence alignment is a way of arranging two or more sequences of DNA in order to identify regions of similarity between the sequences that might signal functional or evolutionary relationships, such as motifs that code proteins (Mount, 2004; Needleman & Wunsch, 1970; Notredame, 2007). Without going into great detail, string matching with errors operates basically by computing the minimal distance between two strings, for instance, by computing the minimum number of basic operations (character insertions, deletions, substitutions) required to transform x into y (for an introduction to string matching, see Duda, Hart, & Stork, 2001, who give an example of how to compute the distance between "excused" and "exhausted" when only letter insertions and substitutions are considered).

We used the MATLAB® nwalign function, with its default parameters (using an algorithm based on the method of Needleman & Wunsch, 1970). This function calculates a global alignment by a form of optimisation that implies that the alignment spans the entire length of the two query sequences.[2] The two required pieces of input are simply the stimulus and the associated response. To make the algorithm work, letters or digits can easily be randomly assigned a codon, which usually specifies one amino acid, itself represented by a letter (e.g., *A* for Alanine, *C* for Cysteine, etc.). A simple function was used to reformat the digits and letters into amino acid letter codes, before entering them into the nwalign function. A second function was used to get the original symbols back. Alignments can be represented in text format. Pipe symbols and colon symbols can be used to indicate identity between two items and substitution between two items, respectively. Figure 2 below demonstrates how alignments are produced by SAA. Every mismatch between two symbols was considered as a mutation /substitution.[3] The nwalign function was used instead of a simpler Levenshtein distance (i.e., the minimum number of operations needed to transform one string into another, with the allowable operations being deletion, substitution, or insertion) in order to adjust the cost of the different operations more freely.

A major advantage in using SAA is that the correspondence between the items and the positions is maintained after a participant makes an error. For instance, after a participant responds "*aemfaeh*" instead of "aemfah", an SAA aligns the *h* letters (see Figure 2 below, example 2) and finds an insertion of the letter "e". This allows the last item to be scored as correctly recalled. Too simple an algorithm for scoring data would indicate systematic position errors after an error is made by a participant, which would tend to overestimate the number of position errors. This point is central in our demonstration. Such difficulties can be avoided by using material drawn without replacement, although this solution might slightly favour recall and limit the study of interference processes, as noted above. For instance, when "*aemh*" is recalled instead of "aemfh", one could merely conclude that the 5th stimuli was correctly recalled. However, using material drawn with replacement, searching for items or counting items in the stimulus, the response can be questionable. For instance, for "*125137*" instead of "12156137", counting the number of 1s reveals that one of them was deleted, but is not possible to determine which "1" item was deleted without thoroughly examining the context in which the "1" items were recalled.

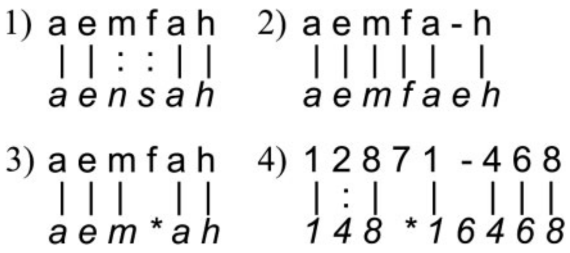

Figure 2. Illustrations of alignments showing 1) a substitution errors 2) an insertion error 3) a deletion error, and 4) a combination of errors. For each of the four examples: Top row = stimulus. Bottom row = participant's response. Middle row = global alignment. Pipe symbol = alignment. Colon = substitution. A dash denotes an insertion in the participant's response. A star denotes a deletion in the participant's response.

Reanalysis of Experiment 1 and Experiment 2 using SAA

We first ran an SAA on both datasets, in order to compute the mean number of items correctly recalled given list length, irrespective of response accuracy (see Figure 3 below). The numbers were computed by summing the number of pipe symbols in every alignment produced by the algorithm. Figure 3 shows a clearly flat performance curve of about 6 items in supraspan conditions (i.e., beyond the person's memory span). The high correlation between the two curves of the mean data points in Figure 3 ($r = .99$, $N = 8$) denotes a similar limitation in span, that is, a limitation of around 6 or 7 items in conditions where the participants were unable to recall the stimuli perfectly. Regarding

[2] We think that a global alignment can fit serial position functions in which good performance for early and late items are observed (local alignment is more complex and can over-prioritize mid-list items).

[3] In general, the algorithms leave a blank for simple substitutions and produce a substitution (:) symbol for conservative substitutions of amino acids whose side chains have similar biochemical properties, but this distinction was not necessary to explore the fundamental capabilities of the algorithms. We simply considered every mismatch as a substitution.

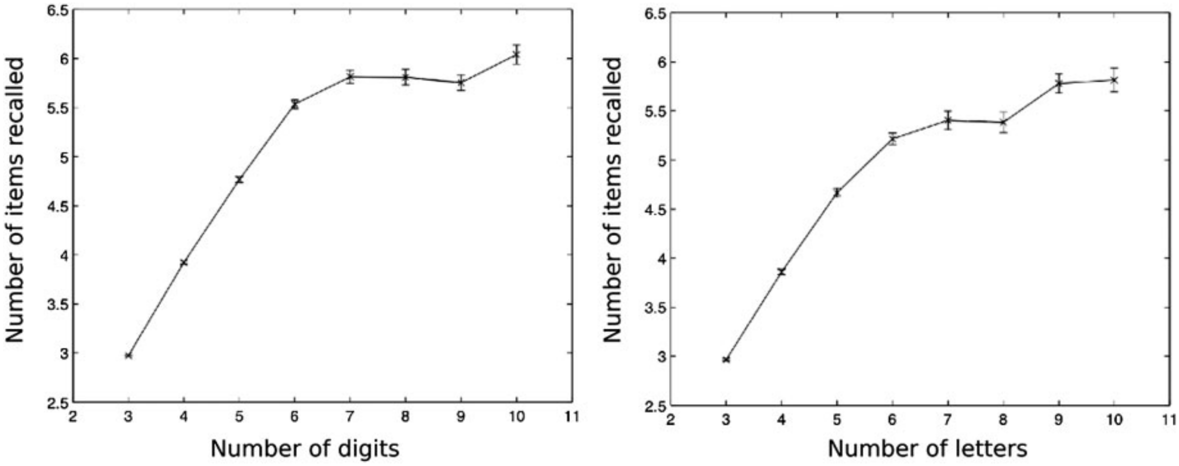

Figure 3. Number of items recalled, by stimulus length, irrespective of response accuracy in Exp. 1 and Exp. 2. Error bars indicate ± one standard error.

the flatness of the curves, two ANOVA (one on lists of 7 digits or more and the other on lists of 9 letters or more) revealed no main effect of stimulus length on the number of items recalled. Also, two inverse functions were fitted to the mean data points and both revealed an asymptote between 6 and 7 items, $R^2 = .99$, $N = 8$, $F > 448$, $p < .001$ for digits, and $R^2 = .97$, $N = 8$, $F > 198$, $p < .001$ for letters. An extrapolation for lists of length 15 gave a mean span of 6.3 digits and 6.6 letters.[4]

We then summed the number of insertion, deletion, and substitution errors for each stimulus, on incorrect responses. Figure 4 below indicates that participants were prone to an exponential number of errors as a function of list length. The mean data points for digits and letters fit by quadratic functions ($R^2 = .99$, MSE = .007, RMSE = .084, and $R^2 = .99$, MSE = .002, RMSE = .049, respectively) were also highly correlated ($r = .99$, $N = 8$). Figure 4 gives a precise idea of the expected number of errors made by participants in cases where they could not recall a stimulus list. For instance, for a list of 7 letters that was not correctly recalled, participants made on average 2.5 errors. Both curves indicate a culmination point of approximately 4.5 errors for lists made up of 10 symbols. Again, the number of items that could be repeated consistently without errors was 3 or 4 when considering the entire set of participants' responses, but this curve still indicates that although few participants made errors in these conditions, those who did made many (around 1.5 errors for a 3- or 4-item list).

Figure 5 below shows how many deletions or substitutions occurred for a given position. For both sets of data, we observe culmination points around .5 for deletions and .2 for substitutions. The number of deletions was similar to the number of omissions predicted by the redintegration model (Lewandowsky, 1999, p. 441, Fig. 4), and close to the number of item errors in Page and Norris (1998, p. 768, Fig. 5). Note that Figure 5 gives the cumulative numbers: for instance, the first mean data point in Figure 5 *A* is based on all stimuli, since all stimuli had one item on position 1. When the number of errors was computed for each stimulus and then divided by the number of stimuli, we observed 84% for deletions of digits (simply meaning that 84 deletions were found for 100 stimuli) and 100% for deletions of letters, and 68% for substitutions of digits and 68% for substitutions of letters, that is, there were approximately 1.5 times as many deletions as substitutions. The percentage of insertions was the smallest (9% for digits and 8% for letters). These differences are comparable to those obtained by McCormack et al. (2000, p. 229), although their proportions were lower since none of their stimulus lengths exceeded six letters. These numbers seem high, but errors did not appear independently and for each stimulus, which explains why they seem above the approximate 50% error rate found in Table 1 below. But still, they indicate an interestingly high

[4] These extrapolations are only given to indicate where performance is supposed to be asymptotic, given the fit, but there is a possibility that lengthier lists would in reality worsen performance and make the task resemble a running-memory span task.

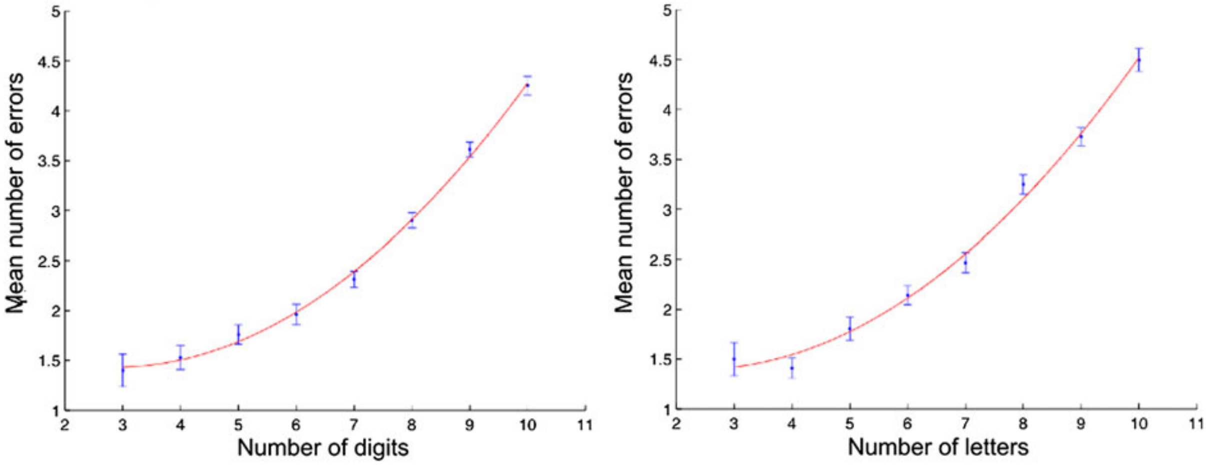

Figure 4. Increase in combined substitution, insertion, and deletion errors, by stimulus length in Exp. 1 and Exp. 2. The graphs indicate the mean number of errors for all imperfect responses (i.e., perfect answers were not taken into account). Error bars indicate ± one standard error.

degree of uncertainty inherent to short-term memory processes. Figure 6 below indicates the fanning effect of error (Farrell & Lewandowsky, 2002), which refers to a decrease in performance at a given serial position as list length increases; this effect was hidden in Figure 5.

A gap accounts for the occurrence of insertions and deletions when pairs of correlated elements in two sequences cannot be identified. The cost of creating a gap is called "gap opening". When the algorithm was used with a non-default value for the gap opening ($GapOpenValue = 2$ instead of $GapOpenValue = 8$), which specifies the penalty for opening a gap in the alignment, the results were quite comparable except that more insertions were found (as expected, since a gap opening allows an insertion operation). Setting the gap opening to a low value was necessary for getting the fourth example in Figure 2 to come out as expected. Using a low penalty ($GapOpenValue = 2$), the two 1s could be aligned, along with the deleted "7" item and the inserted "6" item. Using a higher penalty, 7 was considered to have been replaced by 1, and 1 was considered to have been replaced by 6. Using a low value for the gap opening, the percentage of insertions was greater, and this logically gave rise to a greater number of deletions and a smaller number of substitutions. This demonstrates the flexibility allowed by such algorithms. When the number of errors was computed for each stimulus and then divided by the number of stimuli using the non-default value, we observed 99% for deletions of digits and 115% for deletions of letters, and 43% for substitutions of both digits and letters. The proportion of insertions was higher (24% for both digits and letters). This broadly indicates that deletions were twice as likely as substitutions, and likewise for substitutions compared to insertions, a quite different pattern from the one we obtained with the default value. As a result, the mean data points of Figure 3 were highly correlated ($rs = .99$) with those obtained using the non-default value, indicating a similar trend in the data. However, the increase in the number of insertions found changed the capacity estimates that were given in Figure 3. In spite of the high correlation noted above (signalling a similar trend and a capacity plateau), we observed a higher measure of the number of items recalled using $GapOpenValue = 2$ ($t(7) = 3.0$, $p < .02$ for digits, and $t(7) = 3.7$, $p <$ .008 for letters). The numbers are given in Table 2 below, which still indicate a common limitation around 6 items across the four columns.

Let us take a stimulus-response pair observed in our data (61029296, 69250929).

Using $GapOpenValue = 8$, we obtained:

$$\begin{array}{ccccccccc} 6 & - & 1 & 0 & 2 & 9 & 2 & 9 & 6 \\ | & & : & : & : & | & | & | & \\ 6 & 9 & 2 & 5 & 0 & 9 & 2 & 9 & * \end{array}$$

The algorithm uses a single insertion in order to align the "929" blocked sequence. This alignment presents the advantage that adjacent substitutions can be used as a potential location for transposition errors (for instance, items "0" and "2", which could have been swapped by the participant, could easily be recovered).

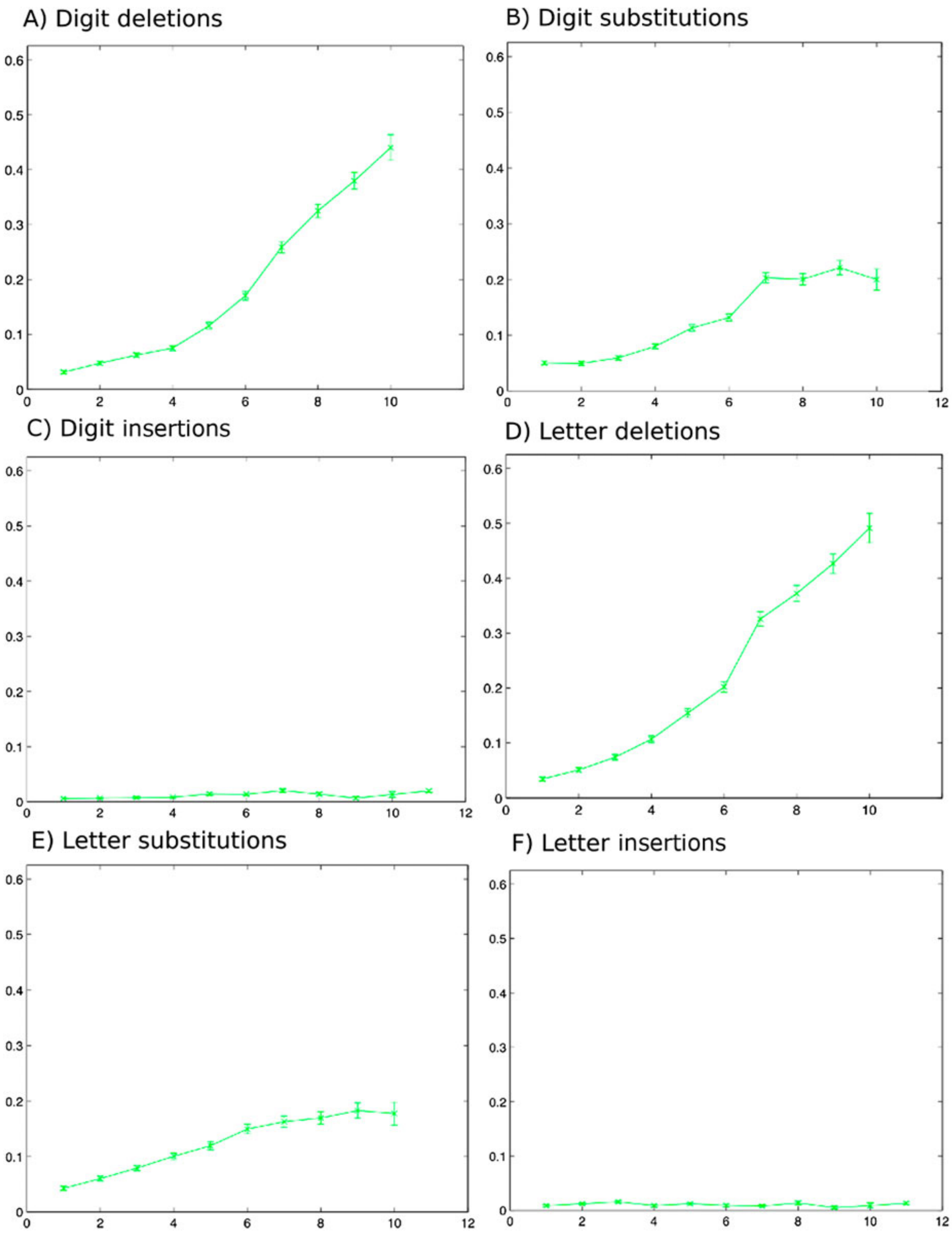

Figure 5. Proportion of deletions, substitutions, and insertion errors by item position in Exp. 1 and Exp. 2. Error bars indicate ± one standard error. Standard errors increased with position number because few stimuli had items with high position numbers whereas a greater number of stimuli had low position number (for instance, all stimuli had items in position 1). In other words, the fanning effect (Farrell & Lewandowsky, 2002), which refers to a decrease in performance at a given serial position as list length increases, is masked here.

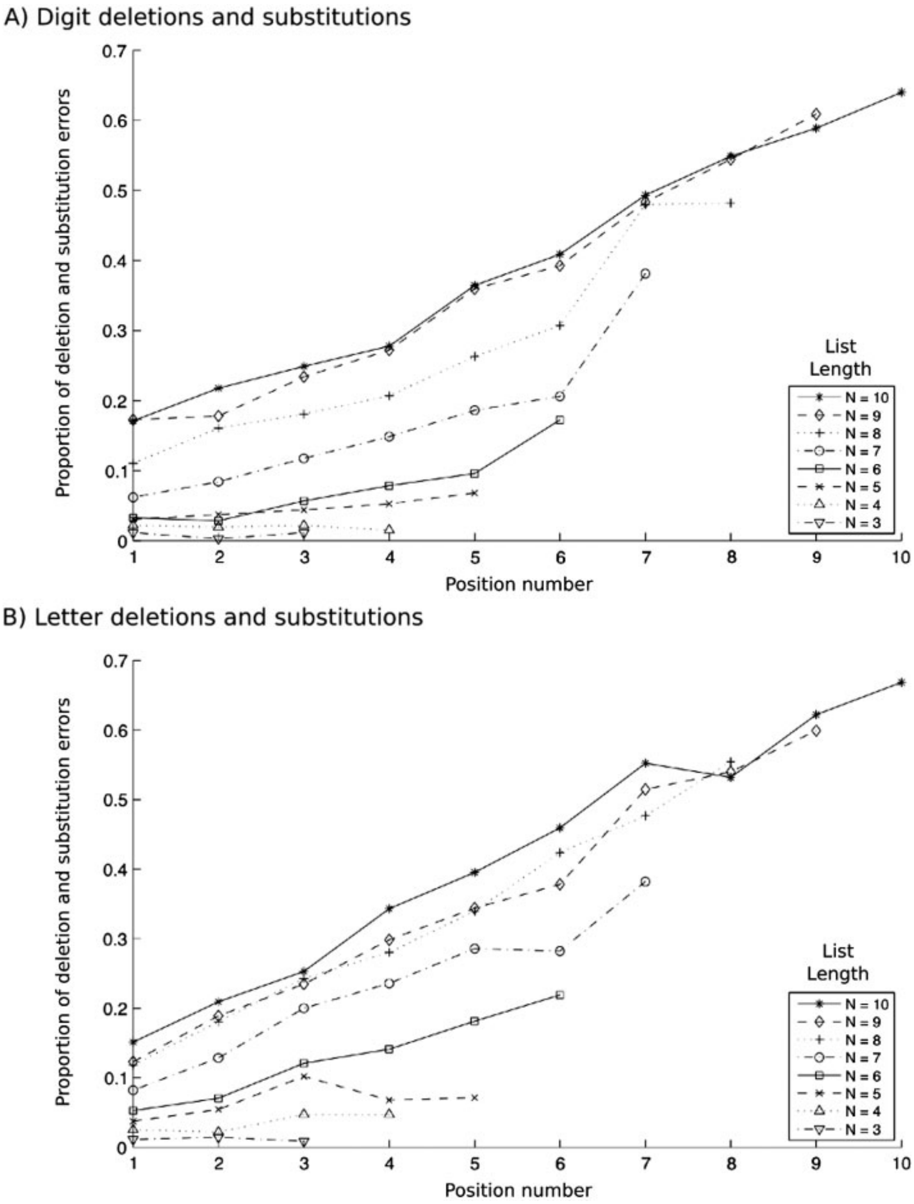

Figure 6. Fanning effect of the proportion of deletions or substitutions (combined), by item position, in Exp. 1 and Exp. 2.

Using *GapOpenValue* = 2, we obtained:

$$\begin{array}{cccccccccc}
6 & - & - & 1 & 0 & 2 & 9 & 2 & 9 & 6 \\
| & & & : & | & & | & | & | & \\
6 & 9 & 2 & 5 & 0 & * & 9 & 2 & 9 & *
\end{array}$$

The algorithm finds one extra insertion and one extra deletion in order to align the "0" item in the middle. Consequently, the number of operations associated with crediting the participant with recalling the "0" item is quite high, so the alignment does not seem plausible.

Using *GapOpenValue* = 1, we obtained:

$$\begin{array}{ccccccccccc}
6 & - & 1 & 0 & 2 & - & - & 9 & 2 & 9 & 6 \\
| & & & & | & & & | & | & | & \\
6 & 9 & * & * & 2 & 5 & 0 & 9 & 2 & 9 & *
\end{array}$$

Now, the alignment considerably increases the number of insertions and deletions in order to align the "2" item in comparison with the *GapOpenValue* = 8 condition. Again, the solution does not seem to reflect plausible psychological errors.

Overall, the default value seemed the most plausible in our data when we scrutinised the different alignments. Future research could profit from individual differences and make use of confirmatory modelling to better parameterise the algorithm, by improving the fit between the manifest variable that would reflect performance on the newly studied task (scored by using a sequence alignment algorithm) and a latent variable estimated from several other tasks. It still remains difficult to predict whether correlations are expected between these tasks and more conventional span tasks.

Our algorithm (available for download) allowed us to conduct an additional analysis because it authorises non-contiguous transposition operations in order to better compute the distance between each stimulus-response pair (e.g., "*[b]nm[a]*" instead of "[a]nm[b]") as well as transpositions of contiguous groups of items (e.g., "*[mb][an]*" instead of "[an][mb]"). Our algorithm produces several error patterns, depending on the cost associated with the transposition error. For instance, given an "868948636" stimulus and a "*989648148*" response made by a participant on the tenth trial, one can notice that the participant recalled a peculiar symmetry at the beginning of the stimulus: "*9896*" instead of "8689".[5] With a very high value associated to a transposition error, the algorithm produces a "*:||-" alignment (deletion of the "8", substitution of "6" with "9", match for "8", match for "9", insertion of "6"). With an intermediate value, the algorithm first transposes "89" with "6" ("*8[89][6]*" instead of "8[6][89]"), and finally finds a substitution of "8" with "9" for the first item, given that "9896" is then compared to "8[89][6]". Finally, using a smaller value, the algorithm transposes the first item with the rest of the sequence in order to indicate that the first item was recalled in the last position (i.e., "[68948636][8]" instead of "[8][68948636]"; then the transposed stimulus "[68948636][8]" can be matched to the participant's response "*989648148*" to get the alignment). A last example: the same participant gave a "6910547839" response for a "6917340537" stimulus (29th trial). In this case, our algorithm found that "691[734][0537]" could be transposed into "691[0537][734]" to better match the response. Thus, we notice that the participant correctly recalled the first three items, "691", then almost correctly recalled "0537" (although the "3" was changed to a "4", and although the block was recalled in the wrong place). For the previous example, the best alignment we obtained was

TABLE 2
Number of items recalled by list length in Exp. 1 and Exp. 2, irrespective of response accuracy, using *GapOpenValue* = 8 (default value; the numbers are plotted in Figure 3) and *GapOpenValue* = 2.

	GapOpenValue =8		*GapOpenValue =2*	
Stim. Length	*Digits*	*Letters*	*Digits*	*Letters*
3	2.97	2.97	2.97	2.97
4	3.92	3.86	3.94	3.87
5	4.77	4.67	4.80	4.70
6	5.53	5.21	5.57	5.29
7	5.81	5.40	5.91	5.53
8	5.81	5.38	5.98	5.56
9	5.75	5.78	5.95	5.96
10	6.04	5.81	6.27	5.99

```
6   [–  2|  1   0]  9   2   9   6
|        |  :   |   |   |   |
6   [9  2|  5   0]  9   2   9   *
```

with a transposition cost equal to 3. In this case, the "2" and "10" sequences are nicely swapped, resulting in the identification of 6 items recalled. However, note that capacity estimates would be different in spirit from those we gave above, which were supposed to reflect immediate recall in correct order. Still, applying this algorithm to the digits, we obtained capacity estimates equal to 6.0, 6.1, 6.1, and 6.5 (for stimulus length ranging from 7 to 10). The numbers are again asymptotic to a capacity of around 6 items, and match the values of Table 2 above.

EXPERIMENT 3 (REPORTED IN MATHY & FELDMAN, 2012)

The first two experiments used simple span tasks that required immediate serial recall of lists in which the frequency of repeated items was quite limited. The new method proved to be adequate for the identification and the quantification of memorisation errors, but the two span tasks that served as our benchmark for testing our method cannot be considered as fundamentally novel. The goal of the last two experiments was to test our method on more novel tasks. We chose to induce

[5] The other peculiar feature is that even though the response hardly matched the stimulus at all, the number of items recalled was correct.

a chunking process by allowing associations to be made between items that presented many regularities (Exp. 3) or repetitions (Exp. 4). These highly patterned lists invited the participants to group and recode the list into chunks. It is well established that the limit appears to be about four chunks when experimental conditions prevent both chunking and rehearsal of the presented items (Cowan, 2001), but Experiment 3 and Experiment 4 aimed at assessing the span in conditions favouring chunking and rehearsal.

In a famous paper, Miller (1956) suggested that the capacity of short-term memory is limited to a "magical number" of about 7 (plus or minus 2) items. More recent research now appears to account for a smaller estimate of about 4 items (Baddeley & Hitch, 1974; Brady, Konkle, & Alvarez, 2009; Broadbent, 1975; Cowan, 2001, 2010; Estes, 1972; Gobet & Clarkson, 2004; Halford, Baker, McCredden, & Bain, 2005; Halford, Wilson, & Phillips, 1998; Luck & Vogel, 1997; Pylyshyn & Storm, 1988; Zhang & Luck, 2008).

In contrast to many memory span tasks in which chunking is deliberately suppressed to achieve correct estimation of working memory capacity (Cowan, 2001), recent research encourages chunk formation processes and gives preference to more liberal short-term memory tasks (introducing redundancies or regularities). These tasks are especially useful to assess capacity in contexts in which associations can be made by participants (Brady et al., 2009; Chen & Cowan, 2005; Mathy & Feldman, 2012). For instance, the bridge between the two previously mentioned estimates (4 ± 1 vs. 7 ± 2) can be made by measuring how much information can be compressed during the task, the true capacity limit appearing to be about 4, while 7 reflecting the number of unpacked items (Mathy & Feldman, 2012).

Method

This section aims at analysing previous data reported in an experiment by Mathy and Feldman (2012). In this experiment, the authors devised a chunking memory span task in which chunking was deliberately facilitated by introducing sequential patterns into the digit sequences. It was hypothesised that the capacity limit is 4 ± 1 "chunks" of information, consistent with many studies (Cowan, 2001). The 23 participants were given an immediate serial list recall task in which lists of digits were created by using increasing or decreasing series of digits (called runs) of variable lengths and increments. The increments were variable between runs within lists, but the increments were constant within runs. For instance, three runs (say, 123, 2468, 43) were concatenated to form a 123246843 list. The length of the sequence was randomly drawn between 3 and 10, the length of the runs was randomly drawn between 1 and 10, and the increment (respectively $+1$, $+2$, and -1 for the previous example) was randomly drawn between -3 and $+3$ (except 0). It was hypothesised that because each of the runs could be chunked, the participant's measured digit span should consistently depend on the number of distinct chunks per sequence, with a limitation of 4 chunks independent of the number of digits that could be unpacked.

In their experiment, all the digits of a given sequence were displayed at once in order to facilitate the chunking process (a procedure inspired by O'Shea & Clegg, 2006). The amount of time during which the entire sequence was shown on the screen was proportional to the number of digits (one second per digit). Whenever a chunk was built by the program, the digits were located on a new line. Therefore, the number of lines simply reflected the number of chunks in a list. The participants were instructed that each line would correspond to a regularity found by the computer. The experimental setting was therefore optimised in order to simplify the memorisation process in that the participants could use both verbal and visual information to chunk the digits that were displayed at once. The participants were instructed to recall the digits in order.

All the other aspects of the experimental procedure were similar to the procedure described in Experiments 1 and 2 of the present paper (timing, number of trials, size of stimuli, response entered on the keyboard, and unpredictable list length).

Results

The chunk span was equal to 4 by integrating under the performance curve (a method similar to the all-or-nothing scoring of Conway et al., 2005). Also, the rate of decline in performance showed that the participants' performance fell below 50% at about 4 chunks, confirming the prediction of Mathy and Feldman (2012). However, these measures do not indicate the number of chunks present in the recall output for each list.

The alignment between the stimulus and the response in the data was therefore computed in order to estimate the actual number of chunks that were both correctly encoded and correctly recalled for each stimulus, irrespective of response accuracy. A correct order of the chunks was taken into account to score performance. The computation of the alignment is particularly helpful in avoiding the incorrect assignment of correct recall to one chunk (false alarm) and avoiding failing to credit a chunk not strictly recalled at the expected position but nevertheless entirely recalled. For instance, given a "123.2.432" stimulus, and a "*123.432*" response, the alignment "||| * |||" signals the omission of the "2" digit in the middle of the sequence, the correct recall of "123" at correct position, and the correct recall of "432" at a lower position than expected. The recall of the chunks was computed in order, beginning with "123", and so on. Once "123" was credited, it could be removed from the subsequent search function. Then, when searching for the "2" one-digit chunk, the algorithm encountered an omission symbol which was associated with incorrect recall. Because the search function followed the alignment, the "2" digit could not be wrongly associated with the one present in the "432" chunk. The resulting estimation of the number of correctly recalled chunks for this example is 2. In sum, the alignment simply allows scoring a chunk as correctly recalled, no matter the strict position of the digits recalled in the participant's response. Figure 7 below indicates the number of chunks that were actually encoded as a function of the number of chunks in the stimulus. The figure shows a logarithmic performance asymptotic to about 4 chunks, $R^2 = .99 (N = 7)$, $p < .001$, $y = .838 + 1.7 \times ln(x)$.

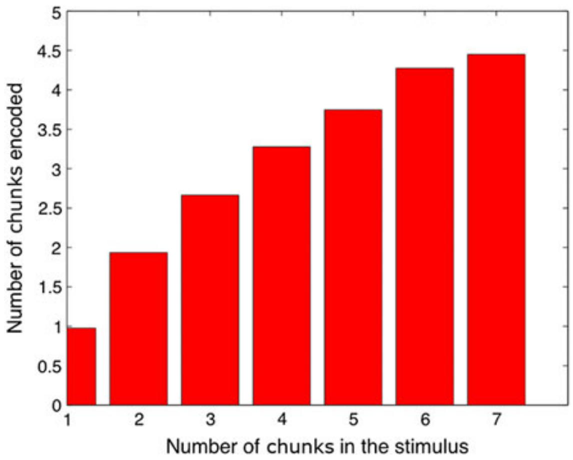

Figure 7. Number of encoded chunks as a function of the number of chunks that were built to form a stimulus in Exp. 3. Adapted from Mathy and Feldman (2012).

The estimation of capacity in this experiment therefore relied on three different scoring criteria that converged to a single value around 4. One advantage of our alignment method is that it allows the manipulation of chunks of various sizes, in contrast to other studies in which chunking is restricted to predictable pairs or triplets of items (Burtis, 1982; Chen & Cowan, 2009; Cowan, Chen, & Rouder, 2004; De Kleine & Verwey, 2009; Li, Blair, & Chow, 2010; O'Shea & Clegg, 2006).

EXPERIMENT 4: INDIVIDUAL DIFFERENCES

Common working memory span tasks such as the reading span (Daneman & Carpenter, 1980) have been thought to measure the ability to simultaneously process and store information. However, as already mentioned in the present paper, the design of such tasks is meant to separate the storage of the to-be-recalled items and the processing of the not-to-be-recalled items. Unsworth et al. (2009) attempted to better understand why such a complex span task correlates so well with measures of higher-order cognition by examining the respective roles of processing and storage using batteries of short-term and working memory span tasks. They found that processing did not fully account for the predictive power of complex span tasks and that simple span tasks were equally predictive of higher-order cognition. An issue is that the role of the processing component in predicting intelligence can be undermined when processing is directed toward the not-to-be-recalled items.

There is also accumulating evidence that suggests that executive functions are related to intelligence (Friedman et al., 2006), but the question of how chunking relates to intelligence is less commonly raised. It quickly becomes apparent that inquiring about the relation between short-term memory, working memory, executive functions, chunking and intelligence requires an extensive methodology! As a start-up experiment, our goal here was simply to begin examining the relation between chunking in short-term memory and working memory. The idea was to show that the use of tasks in which the to-be-recalled items need to be simultaneously processed and stored (for instance, while chunking the to-be-recalled items) is a valuable option to studying the role of the processing component, and that in this condition a correlation can be found with working-memory tasks in which the processing of the to-be-recalled items is privileged (for instance, in tasks requiring the updating of the items).

Method

According to St Clair-Thompson and Sykes (2010), one outstanding issue is whether scoring methods influence the predictive ability of working memory tasks. Although studies have shown that short-term memory and working memory measures are closely related (Colom et al., 2006), appropriate scoring methods might help capture the processing component of working memory tasks, which was hypothesised to add predictive power above and beyond what is accounted for by storage in simpler short-term memory tasks (Unsworth et al., 2009). Unsworth and Engle (2006) for instance suggested that complex spans better predict higher-order cognition because the processing component of working memory span tasks is involved in the retrieval of items that are displaced from primary memory. To simplify matters, as a first step, the present experiment only focused on the correlations between a new chunking memory span task and several working memory span tasks.

This last experiment aims at showing the utility of measuring short-term memory capacity without separating the memory contents from the material to be processed. On the contrary, working memory tasks require that the to-be-recalled items be interspersed with other activity unrelated to the retention of the items. We follow the idea that the processing component should provide some index of the capacity of working memory (Unsworth et al., 2009). The authors argue that if the capacity of working memory is able to simultaneously process and store information, then measures of both processing and storage should be examined together in association. We further argue that processing and storage should be examined together in association when both processes are dedicated to the to-be-recalled items. Our hypothesis is that chunking can adequately measure the processing component of working memory.

The chunking-memory span paradigm (Mathy & Feldman, 2012) was again used in the present study with the same goal of deliberately prompting the grouping of memory items. Many lists allowed possible associations between the items to form related units (for instance the list ‘‘BBFFBBB’’ could be retained as 3 chunks instead of 7 independent letters). In this case, the processing component was thought to contribute to the storage of the letters reorganised in chunks. However, some other lists were only composed of unique letters without repetition (for instance, ‘‘BFKJH’’). This enabled the computation of two different measures of short-tem memory capacity within-subjects: a short-tem memory span for the lists that contained no repetition (using both all-or-nothing and partial-credit scoring), and a less standard span based on the number of correct letters recalled after an alignment was computed when the lists contained repetitions. These three measures were then correlated with a working memory battery. We hypothesised that the span computed with the alignment method (based on conditions in which the to-be-recalled letters were processed to form chunks) would predominantly correlate with a memory updating task that involves a similar process in which the processing component is directly involved in the retention of the items (i.e., the digits are processed before being stored). In order to keep the following analysis short, we do not aim to explain the formation of chunks in this experiment contrary to the previous one in which this goal was targeted (not to mention that the identification of the chunks is far more complicated in the present experiment).

Participants. The participants were 69 Franche-Comté University students who received course credit or 10 euros in exchange for their participation. None of them had participated in any of the previous experiments.

Stimuli. The stimuli were the capital letters (‘‘B’’, ‘‘F’’, ‘‘H’’, ‘‘J’’, ‘‘K’’, ‘‘Q’’, ‘‘R’’, ‘‘T’’, ‘‘X’’, and ‘‘Z’’), chosen for having few phonological similarities in French. The stimuli were displayed visually on a computer screen using MATLAB and the Psychophysics Toolbox (Brainard, 1997). Each letter stimulus was about 2 cm tall, presented in the middle of the screen at a pace of one second per item, and printed in black Times font against a grey background. In a given list of letters, each letter replaced the previous one in the same spatial location. Each list was composed of a maximum of 10 letters.

Procedure. Each experimental session lasted half an hour and included 46 separate stimulus lists. The 46 stimuli and their order were randomly drawn for each participant. After the lists were presented, the participants entered their responses on a keyboard. The time for recall was unlimited and the participants were allowed to correct their responses. They were asked to recall as many letters as possible in the correct order. After the participants validated their answers with the space bar, a feedback screen (plain green vs. plain red) indicated whether the recall was correct (i.e., item memory + order memory both correct) for one second before the next list was displayed.

The 46 lists were built as follows: the length (N) varied from 1 to 10 letters, and the number of different letters (n) varied from 2 to N for each length. This generated the following conditions following the N/n format: 1/1, 2/2, 3/2, 3/3, 4/2, 4/3, 4/4, (...), 10/10, which generated 46 conditions.

Once this first task was terminated, the participants were administered the Working Memory Test (WMT) for MATLAB (Lewandowsky et al., 2010), which we translated into French. The WMT includes four heterogeneous tasks: a memory updating (MU) task, an operation-span (OS) task, a sentence-span (SS) task, and a spatial short-term memory (SSTM) task. In the OS task, the participant saw alternating arithmetic operations and to-be-remembered consonants. The participant had to judge the correctness of each equation while retaining the consonants for later serial recall. The SS task worked in a similar way except that the concurrent task was to judge the meaningfulness of simple sentences. In the SSTM task, the participant had to remember the location of dots in a 10 × 10 matrix. In the MU task, the participant was presented with a set of frames (between 3 to 5 across trials) that contained the to-be-remembered digits displayed for 1 sec each. Following encoding, arithmetic operations (from −7 to +7) were shown in the frames and had to be applied to the digit that was currently remembered in the corresponding frame. The result of the operation on the digit had to replace the preceding memorised digit. Between two and six updates had to be made for each trial, but not every frame was necessarily updated within a trial. The participant typed the updated digits in each frame when prompted by a question mark.

The MU, OS, and SS task capacities were scored as the proportion of items correctly recalled. The authors of the battery make clear that their procedure matches a partial credit scoring system. For instance, in the OS and SS tasks, items were scored as correct when recalled in the correct list position. For the SSTM tasks, performance was computed by awarding 2 points whenever a dot was perfectly recalled, 1 point for a deviation of one cell, and 0 points otherwise.

Again, it was hypothesised that the span computed with the alignment method on conditions in which the participant was induced to form chunks would better correlate with the memory updating task that involved the to-be-recalled digits to be processed before being stored. Our scoring method was not applied to the working-memory-test battery, which computes its own estimation of the span.

Results

To summarise quickly how performance varied with list length (N) and chunking opportunities (n, with less chunking opportunities offered with higher n), we started by computing a linear regression across trials using correct response (scored 1/0 for each trial) as the dependent variable. The result showed that the two factors were significantly predictive of performance (respectively $\beta_N = -.57$ and $\beta_n = -.15$, with an overall $R^2 = .44$), with each increase in N or n resulting in lower performance.

The WMT data were processed using scripts in R (R Development Core Team, 2005) included in the WMT package. The means and SDs for the MU, OS, SS, and SSTM tasks were respectively .50, .64, .61, and .81. These values are slightly lower than those observed by Lewandowsky et al. (2010), but they consistently matched the rather low mean letter spans that we observed in our task (about 5.5, according to the analysis below). The chunking-memory span data were separated into two different data sets: a condition in which there was no repetition in the material and a condition in which repetitions were present in the material. For the first set, a span was computed using both the all-or-nothing and partial-credit scoring methods across the 690 selected trials. For the second data set, we again used the simple MATLAB® nwalign function with its default parameters to compute the number of letters in the response that were correctly aligned with the original list (across the remaining 2484 trials). We simply summed the number of letters in the response that were correctly aligned with the original list to produce a measure of span. Table 3 below shows the correlations obtained between the different methods and the WMT tasks. Confirming our hypothesis, the correlation was highest between the MU task and the span computed with the nwalign method on lists that contained repetitions of letters. The partial-credit method gave better correlations than the all-or-nothing method. Furthermore, the differences between the correlation coefficients, between the partial-credit method and the nwalign one within each column (controlling for the .639 correlation coefficient between partial-credit and nwalign spans) were all significant (all *ts*(66) > 2.37 considering that the *rs* were dependent and using the formula provided by Cohen & Cohen, 1983, p. 57 with $n - 3$ degrees of freedom).

We then used our own alignment method to inquire about transposition errors. The two alignment methods were highly correlated ($r = .95$, $N = 69$). The participants were then split around the median span, computed using the alignment method to distinguish a low-span group from a high-span group. One interesting point is that only the number of deletion errors distinguished our low-span sample from our high-span sample. The mean number of alignments per list separated the two groups significantly (5.2 vs. 6.0, $t(67) = -10$, $p < .001$), not surprisingly because this variable was used to split the sample into two groups, but this was also the case for the number of deletions (1.5 vs. 0.8, $t(67) = 7.5$, $p < .001$). However, the number of substitutions, transpositions and insertions were not significantly different between the groups (the means across both groups were respectively .45, .23, and .19 for the respective errors). Figure 8 below shows the repartition of error types per group. This tends to show that the number of deletions represents most of the variance across groups, as opposed to the other types of error that can be considered more marginal.

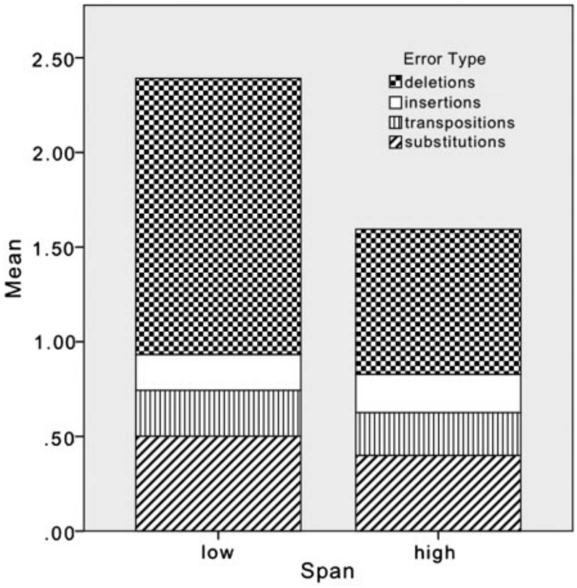

Figure 8. Mean number of deletion, insertion, transposition and substitution errors that were observed according to group (low vs high span) in Exp. 4.

DISCUSSION AND CONCLUSION

Although discussions about immediate serial recall during the last century mainly focused on a subtle examination of the mechanisms underlying memory encoding and retrieval (Nairne, 2002), the scoring of performance is regularly called into question (Blankenship, 1938, p. 10; Conway et al., 2005; Friedman & Miyake, 2005; St Clair-Thompson & Sykes, 2010; Unsworth & Engle, 2007). Effectively, microscopic memory errors need to be accurate in order to test the models. The expression "Garbage in, garbage out" is sometimes used to call attention to the fact that nonsensical data will not be questioned by a model. The aim of this paper was twofold: to study memory for serial order using complex material and to propose an adequate method for scoring performance in this more complicated context.

Presenting a limited number of items with no repeats has many advantages when it comes to controlling the factors underlying performance (e.g., McCormack et al., 2000), but such material may facilitate the recall process, especially when using lists of similar length across trials. We also suspect that simple material is used precisely to facilitate scoring. The cost of using too overly simple material is that participants can find many strategies, for instance to guess what items are in the last few positions. In addition, the limitation of item repetition prevents researchers from discovering remarkable phenomena such as the Ranschburg effect (Henson, 1998a).[6] There is no doubt that there are many other effects to be discovered (Fischer-Baum, 2012).

Another remarkable phenomenon is that secondary memory (St Clair-Thompson & Sykes,

[6] Ranschburg effects in our experiments were naturally favoured since items were sometimes repeated, but our goal was not to detail this effect.

TABLE 3
Correlations between two span measures (All-or-nothing vs. Alignment) and the Working Memory Test tasks in Exp. 4.

Set	*Method*	*Mean span* (SD)	*MU*	*OS*	*SS*	*SSTM*
no-repetition	All-or-nothing	5.4 (1.3)	.319**	.278*	.157	.142
no-repetition	Partial-credit	5.4 (1.3)	.336**	.420**	.249*	.191
repetitions	nwalign	5.6 (0.6)	.631**	.531**	.395**	.245*

**p* <.05; ***p* <.01; N = 69; MU = memory updating task; OS = operation-span task; SS = sentence-span task; SSTM = spatial short-term memory task.

2010; Unsworth & Engle, 2007) might be involved to a greater extent when span tasks involve greater processing of the to-be-recalled items. This is the case when repetitive information can be chunked (Exp. 3 and Exp. 4). In turn, such tasks require the use of more complex scoring methods able to capture performance. The aim of the present study was to develop such a scoring method, based on the use of sequence alignment algorithms. In light of our results, we highly recommend this method that facilitates the identification of many usual memory errors, and which eventually produces an estimate of the span that is not underestimated. One interesting example is that the span seems to remain constant in supraspan conditions (about 6 digits or letters in Exp. 1 and Exp. 2, and about 4 chunks in Exp. 3), as we observed stabilised performance (the slight plateaus in Figures 3 and 7 tend to indicate that there is no deterioration of the encoded items when capacity is exceeded). We believe that this method makes a significant contribution to the short-term memory domain, as it can provide usual descriptions of the data, such as fanning effects (Figure 6), as well as less usual descriptions such as the quantification of memory errors with list length (Figure 4) and the repartitioning of memory errors in low versus high span groups (Exp. 4). Finally, we showed that interesting relationships can be found between different estimates and common working memory tests (Exp. 4).

We also agree with Unsworth and Engle (2007) that a particular process may affect performance more on one task than on the other. For instance, the chunking memory span tasks used in Exp. 3 and Exp. 4 directly test the capacity of short-term memory to simultaneously process and store information. The chunking aspect of the tasks suggests that processing efficiency and accuracy should be positively correlated: the participants who are more efficient at processing better encode information and subsequently obtain higher span scores. We believe that there is a dire need to extend experimentation on immediate serial recall to such complex material. For instance, Unsworth et al. (2009) argue that the processing component is often overlooked in complex span tasks in which relatively few errors are committed, and that other studies which have attempted to use single-task (processing-only) performance seem ill-suited to learning about how processing performance affects recall performance during the more specific complex span tasks. Our tasks present another potential advantage. We believe that developing more complex tasks (with adequate scoring methods) such as those presented in this paper could be useful for increasing similarity-based interference errors to further test models of short-term memory capacity limits (e.g., Oberauer & Kliegl, 2006).

Concerning the alignment method per se, Friedman and Miyake (2005) recommended that researchers score span tasks with continuous measures (i.e., partial-credit rather than all-or-nothing), a point of view supported by our results of Exp. 4. We believe that the sequence alignment method that we used to score performance in all of our experiments directly follows this suggestion, because it allows the estimation of a proportion of correct responses in the case that many errors are committed. In addition, offering a new method for scoring performance can also help in designing experiments. For instance, Chen and Cowan (2005, 2009) asked participants to leave a slot blank if they forgot a word for a particular serial position (see also McCormack et al., 2000, who used the same procedure). Our scoring method would have been useful in this case through letting participants report the items they could remember more naturally.

The present study demonstrates the benefits of the serial-recall paradigm, especially when rehearsal and grouping are encouraged. Working-memory tasks implying a concurrent task have their own qualities, but there is potential for other memory tasks. For instance, we found that a reliable estimator of individual differences in working memory tasks could be measured using complex material with repeatable patterns (Exp. 4). We showed that estimations of digit and letter spans can be computed (Exp. 1 and Exp. 2), as well as estimations of capacity in terms of chunks (Exp. 3) in supraspan conditions where many errors occur in memory. The following assumptions were made: (1) the types of operations are considered fundamental for comparing a stimulus and a response (insertions, omissions, substitutions, etc.) as well as the cost of each operation (if a substitution is thought to be less probable than an omission, it needs to be associated with a higher cost); (2) the value of the parameter for allowing the intrusions of items must be determined (some prototypical examples can be used to choose the best options); (3) a global or local approach is necessary for analysing the sequences, depending on whether long sequences of elements are entirely forgotten by participants; and (4) a permutation operation must be included in the algorithm. To return to our introductory examples, when the nwalign function is run with its default value (i.e., with $GapOpenValue = 8$), the alignment proposed for the fourth example in Figure 2 is "|:|::|||", meaning that the "7" and "1" symbols were replaced by "1" and "6", respectively. It is only when $GapOpenValue = 2$ that the alignment shown in Figure 2 is produced, with the "1" correctly aligned. However, it is difficult to believe that one or the other of the two alignments best characterises the errors that occur in memory for this particular example. Such parameterisation flexibility can be profitable for model performance in various populations, in particular for examining developmental changes in serial-recall error patterns (Maylor, Vousden, & Brown, 1999; McCormack et al., 2000). We believe that this flexibility does not undermine the fact that alignments can be computed easily and rapidly for both serial data with or without repeatable patterns. Finally, the most important question is not to know if the estimation is irreproachable, but rather to know if the memorisation lists which contain repetitions is a worthwhile undertaking in the study of memory processes. If the usage of lists with repetitions is worthwhile (e.g., Hu & Ericsson, 2012) then it is necessary to find a way to evaluate performance.

A couple of problems should be addressed in future research. In relation to what was stated above, although an estimate of short-term memory span can be made by running a sequence alignment algorithm, the estimate is subject to slight variations that depend on how the algorithm is set up. Calibration can be done, for instance, by checking the alignments on a few critical examples before testing the algorithm on a larger data set. Another problem is that searching for items that have been moved several positions away (Henson, 1996) requires additional analysis. For instance, if instead of recalling "abcde", a participant recalls "*bcdea*", the algorithm might indicate that the first "a" item of the sequence was deleted and that the last "a" item of the response was inserted. Further computation is needed in order to check whether the deleted and inserted items are the same, in which case the translocation error can be identified if it was not detected during the initial analysis. A final problem pertaining to the choices, we made concerning the experiments is that our experiments used unpredictable list lengths, which is known to limit capacity in comparison to fixed-length list tasks (Bunting, Cowan, & Colflesh, 2008). There is a possibility that this choice also reduced the stability in the patterns of errors we observed.

SAA offers many options that could prompt further studies. For instance, subsequent research would benefit from using the same set of stimulus sequences across participants. The use of multiple sequence alignment is an idea that needs to be pursued in order to characterise inter-participant error patterns and their variance. A multiple alignment can show all the sequences along a consensus sequence that is determined by the alignment. This method would tend to reinforce the single analysis produced by simple sequence alignments by assigning a given type of error to a given position with greater certainty. The consensus sequence would guarantee that the best alignment is found before the error patterns are analysed. In addition, by recoding several stimuli presented sequentially to participants, by position (for instance, "ahfc" and "rtes" both be represented as "1234"), the algorithm could be used for finding proactive interferences of positional information (or protrusions) in other words, for finding the likelihood that an erroneous item in one trial would occur in the same position in the previous trial (Henson, 1996, 1998b).

This study focused on a method revisited by bioinformatics aimed at measuring performance and providing the best possible characterisation of retention-error patterns in complex immediate serial recall tasks. We believe that the systematic error patterns that such a method may reveal will help shed light on the mechanisms underlying both remembering and forgetting.

Manuscript received 14 December 2012
Manuscript accepted 16 December 2012
First published online 12 March 2013

REFERENCES

Aben, B., Stapert, S., & Blokland, A. (2012). About the distinction between working memory and short-term memory. *Frontiers in Psychology*, *3*(301). doi: 10.3389/fpsyg.2012.00301

Anderson, J. R., & Matessa, M. (1997). A production system theory of serial memory. *Psychological Review*, *104*, 728–748.

Baddeley, A. (1986). *Working memory*. New York, NY: Oxford University Press.

Baddeley, A., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *Recent advances in learning and motivation* (pp. 647–667). New York, NY: Academic Press.

Barouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General*, *133*, 83–100.

Barrouillet, P., & Camos, V. (2012). As time goes by: Temporal constraints in working memory. *Current Directions in Psychological Science*, *21*, 413–419.

Blankenship, A. B. (1938). Memory span: A review of the literature. *Psychological Bulletin*, *35*, 1–25.

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review*, *113*, 201–233.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*, 487–502.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Broadbent, D. (1975). The magic number seven after fifteen years. In A. Kennedy & A. Wilkes (Eds.), *Studies in long-term memory* (pp. 3–18). New York, NY: Wiley.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*, 539–576.

Brown, G. D. A., Preece, T., & Hulme, C. (2000). Oscillator-based memory for serial order. *Psychological Review*, *107*, 127–181.

Bunting, M., Cowan, N., & Colflesh, G. H. (2008). The deployment of attention in short-term memory tasks: Trade-offs between immediate and delayed deployment. *Memory & Cognition*, *36*, 799–812.

Bunting, M., Cowan, N., & Saults, J. S. (2006). How does running memory span work? *Quarterly Journal of Experimental Psychology*, *59*, 1691–1700.

Burgess, N., & Hitch, G. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, *106*, 551–581.

Burtis, P. J. (1982). Capacity increase and chunking in the development of short-term memory. *Journal of Experimental Child Psychology*, *34*, 387–413.

Chekaf, M., & Mathy, F. (2012). *Chunking memory span of categorizable objects*. Paper presented at the 53rd Annual Meeting of the Psychonomic Society. 15–18 November, Minneapolis, MN.

Chekaf, M., & Mathy, F. (in revision). Chunking memory span of categorizable objects.

Chen, Z., & Cowan, N. (2005). Chunk limits and length limits in immediate recall: A reconciliation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1235–1249.

Chen, Z., & Cowan, N. (2009). Core verbal working-memory capacity: The limit in words retained without covert articulation. *Quarterly Journal of Experimental Psychology*, *62*, 1420–1429.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum.

Colom, R., Rebollo, I., Abad, F. J., & Shih, P. C. (2006). Complex span tasks, simple span tasks, and cognitive abilities: A reanalysis of key studies. *Memory & Cognition*, *34*, 158–171.

Conrad, R. (1964). Acoustic confusion in immediate memory. *British Journal of Psychology*, *55*, 75–84.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*, 769–786.

Conway, A. R., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences 7*, 547–552.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*, 51–57.

Cowan, N., Chen, Z., & Rouder, J. N. (2004). Constant capacity in an immediate serial-recall task: A logical sequel to miller (1956). *Psychological Science*, *15*, 634–640.

Crannell, C. W., & Parrish, J. M. (1957). A comparison of immediate memory span for digits, letters, and words. *Journal of Psychology: Interdisciplinary and Applied*, *44*, 319–327.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior*, *19*, 450–466.

de Abreu, P. M. E., Conway, A. R., & Gathercole, S. E. (2010). Working memory and fluid intelligence in young children. *Intelligence*, *38*, 552–561.

De Kleine, E., & Verwey, W. B. (2009). Motor learning and chunking in dyslexia. *Journal of Motor Behavior*, *41*, 331–337.

Della Sala, S., Gray, C., Baddeley, A., Allamano, N., & Wilson, L. (1999). Pattern span: A tool for unwelding visuo-spatial memory. *Neuropsychologia*, *37*, 1189–1199.

Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification*. New York, NY: John Wiley & Sons.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*, 19–23.

Estes, W. K. (1972). An associative basis for coding and organization in memory. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory* (pp. 161–190). Washington, DC: Winston.

Estes, W. K. (1997). Processes of memory loss, recovery, and distortion. *Psychological Review*, *104*, 148–169.

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review*, *9*, 59–79.

Fischer-Baum, S. (2012). *Repetition schemas in immediate serial recall*. Paper presented at the 53rd Annual Meeting of the Psychonomic Society, 15–18 November, Minneapolis, MN.

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*, 614–636.

Friedman, N. P., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior Research Methods*, *37*, 581–590.

Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., Defries, J. C., & Hewitt, J. K. (2006). Not all executive functions are related to intelligence. *Psychological Science*, *17*, 172–179.

Gallistel, C. R. (1990). *The organization of learning*. Cambridge, MA: MIT Press.

Gobet, F., & Clarkson, G. (2004). Chunks in expert memory: Evidence for the magical number four … or is it two? *Memory*, *12*, 732–747.

Halford, G. S., Baker, R., McCredden, J. E., & Bain, J. D. (2005). How many variables can humans process? *Psychological Science*, *16*, 70–76.

Halford, G. S., Wilson, W. H., & Phillips, W. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental and cognitive psychology. *Behavioral and Brain Sciences*, *21*, 803–831.

Healy, A. F. (1974). Separating item from order information in short-term memory. *Journal of Verbal Learning & Verbal Behavior*, *13*, 644–655.

Henson, R. N. A. (1996). *Short-term memory for serial order* (Unpublished doctoral dissertation). University of Cambridge, UK.

Henson, R. N. A. (1998a). Item repetition in short-term memory: Ranschburg repeated. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1162–1181.

Henson, R. N. A. (1998b). Short-term memory for serial order: The start-end model. *Cognitive Psychology*, *36*, 73–137.

Henson, R. N. A., Norris, D. G., Page, M. P. A., & Baddeley, A. D. (1996). Unchained memory: Error patterns rule out chaining models of immediate serial recall. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *49*, 80–115.

Hu, Y., & Ericsson, K. A. (2012). Memorization and recall of very long lists accounted for within the long-term working memory framework. *Cognitive Psychology*, *64*, 235–266.

Lewandowsky, S. (1999). Redintegration and response suppression in serial recall: A dynamic network model. *International Journal of Psychology*, *34*, 434–446.

Lewandowsky, S., & Murdock, B. B. (1989). Memory for serial order. *Psychological Review*, *96*, 25–57.

Lewandowsky, S., Oberauer, K., Yang, L. X., & Ecker, U. K. (2010). A working memory test battery for MATLAB. *Behavior Research Methods*, *42*, 571–585.

Li, K. Z., Blair, M., & Chow, V. S. (2010). Sequential performance in young and older adults: Evidence of chunking and inhibition. *Aging Neuropsycholy, and Cognition*, *17*, 270–295.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281.

Majerus, S., Poncelet, M., Elsen, B., & van der Linden, M. (2006). Exploring the relationship between new word learning and short-term memory for serial order recall, item recall, and item recognition. *European Journal of Cognitive Psychology*, *18*, 848–873.

Martin, M. (1978). Memory span as a measure of individual differences in memory capacity. *Memory & Cognition*, *6*, 194–198.

Mathy, F., & Feldman, J. (2009). A rule-based presentation order facilitates category learning. *Psychonomic Bulletin & Review*, *16*, 1050–1057.

Mathy, F., & Feldman, J. (2012). What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition*, *122*, 346–362.

Maylor, E. A., Vousden, J. I., & Brown, G. D. A. (1999). Adult age differences in short-term memory for serial order: Data and a model. *Psychology and Aging*, *14*, 572–594.

McCormack, T., Brown, G. D. A., Vousden, J. I., & Henson, R. N. A. (2000). Children's serial recall errors: Implications for theories of short-term memory development. *Journal of Experimental Child Psychology*, *76*, 222–252.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Morris, N., & Jones, D. M. (1990). Memory updating in working memory: The role of the central executive. *British Journal of Psychology*, *81*, 111–121.

Mount, D. M. (2004). *Bioinformatics: Sequence and genome analysis* (2nd ed.). Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*, 482–488.

Murdock, B. B. (1995). Developing TODAM: Three models for serial-order information. *Memory & Cognition*, *23*, 631–645.

Nairne, J. S. (1992). The loss of positional certainty in long-term memory. *Psychological Science*, *3*, 199–202.

Nairne, J. S. (2002). Remembering over the short-term: The case against the standard model. *Annual Review of Psychology*, *53*, 53–81.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*, 443–453.

Notredame, C. (2007). Recent evolutions of multiple sequence alignment algorithms. *PLoS Computational Biology*, *3*, e123.

Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of Memory and Language*, *55*, 601–626.

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic Bulletin & Review*, *19*, 779–819.

O'Shea, G., & Clegg, B. A. (2006). Stimulus and response chunking in the Hebb Digits task. *Psychological Research*, *70*, 180–192.

Page, M. P., & Norris, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review*, *105*, 761–781.

Pylyshyn, Z., & Storm, R. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, *3*, 1–19.

R Development Core Team. (2005). *R: A language and environment for statistical computing*. Vienna: R Foundation.

Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 592–608.

St Clair-Thompson, H., & Sykes, S. (2010). Scoring methods and the predictive ability of working memory tasks. *Behavior Research Methods*, *42*, 969–975.

Surprenant, A. M., & Neath, I. (2009). The 9 lives of short-term memory. In: A. Thorn & M. Page (Eds.), *Interactions between short-term and long-term memory in the verbal domain* (pp. 16–43). Hove: Psychology Press.

Unsworth, N., & Engle, R. W. (2006). Simple and complex memory spans and their relation to fluid abilities: Evidence from list-length effects. *Journal of Memory and Language*, *54*, 68–80.

Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin*, *133*, 1038–1066.

Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., & Engle, R. W. (2009). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory*, *17*, 635–654.

Zhang, W., & Luck, S. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*, 233–236.

Copyright of Memory is the property of Psychology Press (UK) and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.